Data Adquisition

Jose Hdz

February 19, 2021

Contents

1 Function declaration

```
import requests
import io
from bs4 import BeautifulSoup
import pandas as pd
from tabulate import tabulate

def get_soup(url: str) -> BeautifulSoup:
    response = requests.get(url)
    return BeautifulSoup(response.content, 'html.parser')

def get_csv_from_url(url:str) -> pd.DataFrame:
    s=requests.get(url).content
    return pd.read_csv(io.StringIO(s.decode('utf-8')))

def print_tabulate(df: pd.DataFrame):
    print(tabulate(df, headers=df.columns, tablefmt='orgtbl'))
```

1.1 UANL

```
soup = get_soup(f"http://transparencia.uanl.mx/remuneraciones_mensuales/bxd.php?pag_act
table = soup.find_all("table")[1].find_all('a')
print(table)
```

```
from typing import List, Tuple

def limpiar_nombre_dependencia(nombre_sucio:str)->str:
    nombre_en_partes = nombre_sucio.split(' ')
    return ' '.join(nombre_en_partes[2:])

def obtener_cantidad_de_filas(df: pd.DataFrame)-> int:
    return len(df.index)

def limpiar_dato_sueldo(sueldo_txt: str)-> float:
    return float(sueldo_txt[2:].replace(",", ""))

def get_dependencias_uanl()-> Tuple[List,List[str],List[str]]:
    soup = get_soup(f"http://transparencia.uanl.mx/remuneraciones_mensuales/bxd.php")
    table = soup.find_all("table")[0].find_all('tr')
    listado_dependencias = [(option['value'], limpiar_nombre_dependencia(option.text))
    listado_meses = [option['value'] for option in table[2].find_all('td')[0].find_all
    listado_anios = [option['value'] for option in table[2].find_all('td')[1].find_all
    return (listado_dependencias,listado_meses, listado_anios)

def get_pages(periodo: str, area: str)-> List[str]:
    soup = get_soup(f"http://transparencia.uanl.mx/remuneraciones_mensuales/bxd.php?pag
    try:
        links = soup.find_all("table")[1].find_all('a')
    except Exception as e:
        print(e)
        return []
    return ['1'] + [link.text for link in links]

def get_info_transparencia_uanl(periodo: str, area: str, page:int = 1) -> pd.DataFrame
    soup = get_soup(f"http://transparencia.uanl.mx/remuneraciones_mensuales/bxd.php?pag
    table = soup.find_all("table")
    try:
        table_row = table[2].find_all('tr')
        list_of_lists = [[row_column.text.strip() for row_column in row.find_all('td')]
```

```
        df = pd.DataFrame(list_of_lists[1:], columns=list_of_lists[0])
        df["Sueldo Neto"] = df["Sueldo Neto"].transform(limpiar_dato_sueldo)
        df = df.drop(['Detalle'], axis=1)
    except Exception as e:
        print(f"pagina sin informacion a: {area}, per: {periodo}, page:{page}")
        print(e)
        df = pd.DataFrame()
    return df

def unir_datos(ldf: List[pd.DataFrame], dependencia:str, mes: str, anio:str) -> pd.Data
    if len(ldf) > 0:
        df = pd.concat(ldf)
        df["dependencia"] = [dependencia for i in range(0, obtener_cantidad_de_filas(d
        df["mes"] = [mes for i in range(0, obtener_cantidad_de_filas(df))]
        df["anio"] = [anio for i in range(0, obtener_cantidad_de_filas(df))]
    else:
        df= pd.DataFrame()
    return df
listado_dependencias, listado_meses, listado_anios = get_dependencias_uanl()
# dependencia = listado_dependencias[1]
# mes = listado_meses[10]
# anio = listado_anios[1]
# print(f"{dependencia[1]} {mes} {anio}")
# pages = get_pages(f"{mes}{anio}", dependencia[0])
# dfs = [get_info_transparencia_uanl(f"{mes}{anio}", dependencia[0], page) for page in
# df = pd.concat(dfs)
# print_tabulate(df)
# df.to_csv("uanl.csv", index=False)


#if len(dfs)> 0:
#    df = pd.concat(dfs)
#print_tabulate(df)
ldfs = []
for anio in listado_anios[1:]:
    for mes in listado_meses:
        for dependencia in listado_dependencias[1:5]:
            pages = get_pages(f"{mes}{anio}", dependencia[0])
            print(f"m: {mes} a: {anio} d: {dependencia} p: {pages}")
            ldf = [get_info_transparencia_uanl(f"{mes}{anio}", dependencia, page) for p
```

```
            udf = unir_datos(ldf, dependencia, mes, anio)
            ldfs.append(udf)
df = pd.concat(ldfs)
df.to_csv("uanl.csv", index=False)
# df1 = get_info_transparencia_uanl(1)
# dfs = [get_info_transparencia_uanl(page) for page in range(1,10)]
# df = pd.concat(dfs)
# print(df)
# df.to_csv("uanl_ini.csv", index=False)
```

m: 01 a: 2020 d: ('1103', 'RECTORIA') p: ['1'] pagina sin informacion a: ('1103', 'RECTORIA'), per: 012020, page:1 list index out of range m: 01 a: 2020 d: ('1104', 'SECRETARIA GENERAL') p: ['1', '2'] pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 012020, page:1 list index out of range pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 012020, page:2 list index out of range m: 01 a: 2020 d: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS') p: ['1', '2', '3'] pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 012020, page:1 list index out of range pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 012020, page:2 list index out of range pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 012020, page:3 list index out of range m: 01 a: 2020 d: ('1202', 'SRIA. DE INVESTIGACIÓN CIENTÍFICA Y DESARROLLO TECNOLÓGICO') p: ['1'] pagina sin informacion a: ('1202', 'SRIA. DE INVESTIGACIÓN CIENTÍFICA Y DESARROLLO TECNOLÓGICO'), per: 012020, page:1 list index out of range m: 02 a: 2020 d: ('1103', 'RECTORIA') p: ['1'] pagina sin informacion a: ('1103', 'RECTORIA'), per: 022020, page:1 list index out of range m: 02 a: 2020 d: ('1104', 'SECRETARIA GENERAL') p: ['1', '2', '3'] pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 022020, page:1 list index out of range pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 022020, page:2 list index out of range pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 022020, page:3 list index out of range m: 02 a: 2020 d: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS') p: ['1', '2', '3'] pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 022020, page:1 list index out of range pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION

Y PROYECTOS ESTRATEGICOS'), per: 022020, page:2 list index out of range pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 022020, page:3 list index out of range m: 02 a: 2020 d: ('1202', 'SRIA. DE INVESTIGACIÓN CIENTÍFICA Y DESARROLLO TECNOLÓGICO') p: ['1'] pagina sin informacion a: ('1202', 'SRIA. DE INVESTIGACIÓN CIENTÍFICA Y DESARROLLO TECNOLÓGICO'), per: 022020, page:1 list index out of range m: 03 a: 2020 d: ('1103', 'RECTORIA') p: ['1'] pagina sin informacion a: ('1103', 'RECTORIA'), per: 032020, page:1 list index out of range m: 03 a: 2020 d: ('1104', 'SECRETARIA GENERAL') p: ['1', '2', '3'] pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 032020, page:1 list index out of range pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 032020, page:2 list index out of range pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 032020, page:3 list index out of range m: 03 a: 2020 d: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS') p: ['1', '2', '3'] pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 032020, page:1 list index out of range pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 032020, page:2 list index out of range pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 032020, page:3 list index out of range m: 03 a: 2020 d: ('1202', 'SRIA. DE INVESTIGACIÓN CIENTÍFICA Y DESARROLLO TECNOLÓGICO') p: ['1'] pagina sin informacion a: ('1202', 'SRIA. DE INVESTIGACIÓN CIENTÍFICA Y DESARROLLO TECNOLÓGICO'), per: 032020, page:1 list index out of range m: 04 a: 2020 d: ('1103', 'RECTORIA') p: ['1'] pagina sin informacion a: ('1103', 'RECTORIA'), per: 042020, page:1 list index out of range m: 04 a: 2020 d: ('1104', 'SECRETARIA GENERAL') p: ['1', '2', '3'] pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 042020, page:1 list index out of range pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 042020, page:2 list index out of range pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 042020, page:3 list index out of range m: 04 a: 2020 d: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS') p: ['1', '2', '3'] pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 042020, page:1 list index out of range pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 042020, page:2 list index out of range pagina sin informacion a:

('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 042020, page:3 list index out of range m: 04 a: 2020 d: ('1202', 'SRIA. DE INVESTIGACIÓN CIENTÍFICA Y DESARROLLO TECNOLÓGICO') p: ['1'] pagina sin informacion a: ('1202', 'SRIA. DE INVESTIGACIÓN CIENTÍFICA Y DESARROLLO TECNOLÓGICO'), per: 042020, page:1 list index out of range m: 05 a: 2020 d: ('1103', 'RECTORIA') p: ['1'] pagina sin informacion a: ('1103', 'RECTORIA'), per: 052020, page:1 list index out of range m: 05 a: 2020 d: ('1104', 'SECRETARIA GENERAL') p: ['1', '2', '3'] pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 052020, page:1 list index out of range pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 052020, page:2 list index out of range pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 052020, page:3 list index out of range m: 05 a: 2020 d: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS') p: ['1', '2', '3'] pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 052020, page:1 list index out of range pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 052020, page:2 list index out of range pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 052020, page:3 list index out of range m: 05 a: 2020 d: ('1202', 'SRIA. DE INVESTIGACIÓN CIENTÍFICA Y DESARROLLO TECNOLÓGICO') p: ['1'] pagina sin informacion a: ('1202', 'SRIA. DE INVESTIGACIÓN CIENTÍFICA Y DESARROLLO TECNOLÓGICO'), per: 052020, page:1 list index out of range m: 06 a: 2020 d: ('1103', 'RECTORIA') p: ['1'] pagina sin informacion a: ('1103', 'RECTORIA'), per: 062020, page:1 list index out of range m: 06 a: 2020 d: ('1104', 'SECRETARIA GENERAL') p: ['1', '2', '3'] pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 062020, page:1 list index out of range pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 062020, page:2 list index out of range pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 062020, page:3 list index out of range m: 06 a: 2020 d: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS') p: ['1', '2', '3'] pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 062020, page:1 list index out of range pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 062020, page:2 list index out of range pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 062020, page:3 list index out of

range m: 06 a: 2020 d: ('1202', 'SRIA. DE INVESTIGACIÓN CIENTÍFICA Y DESARROLLO TECNOLÓGICO') p: ['1'] pagina sin informacion a: ('1202', 'SRIA. DE INVESTIGACIÓN CIENTÍFICA Y DESARROLLO TECNOLÓGICO'), per: 062020, page:1 list index out of range m: 07 a: 2020 d: ('1103', 'RECTORIA') p: ['1'] pagina sin informacion a: ('1103', 'RECTORIA'), per: 072020, page:1 list index out of range m: 07 a: 2020 d: ('1104', 'SECRETARIA GENERAL') p: ['1', '2', '3'] pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 072020, page:1 list index out of range pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 072020, page:2 list index out of range pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 072020, page:3 list index out of range m: 07 a: 2020 d: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS') p: ['1', '2', '3'] pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 072020, page:1 list index out of range pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 072020, page:2 list index out of range pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 072020, page:3 list index out of range m: 07 a: 2020 d: ('1202', 'SRIA. DE INVESTIGACIÓN CIENTÍFICA Y DESARROLLO TECNOLÓGICO') p: ['1'] pagina sin informacion a: ('1202', 'SRIA. DE INVESTIGACIÓN CIENTÍFICA Y DESARROLLO TECNOLÓGICO'), per: 072020, page:1 list index out of range m: 08 a: 2020 d: ('1103', 'RECTORIA') p: ['1'] pagina sin informacion a: ('1103', 'RECTORIA'), per: 082020, page:1 list index out of range m: 08 a: 2020 d: ('1104', 'SECRETARIA GENERAL') p: ['1', '2', '3'] pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 082020, page:1 list index out of range pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 082020, page:2 list index out of range pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 082020, page:3 list index out of range m: 08 a: 2020 d: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS') p: ['1', '2', '3'] pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 082020, page:1 list index out of range pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 082020, page:2 list index out of range pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 082020, page:3 list index out of range m: 08 a: 2020 d: ('1202', 'SRIA. DE INVESTIGACIÓN CIENTÍFICA Y DESARROLLO TECNOLÓGICO') p:

['1'] pagina sin informacion a: ('1202', 'SRIA. DE INVESTIGACIÓN CIENTÍFICA Y DESARROLLO TECNOLÓGICO'), per: 082020, page:1 list index out of range m: 09 a: 2020 d: ('1103', 'RECTORIA') p: ['1', '2'] pagina sin informacion a: ('1103', 'RECTORIA'), per: 092020, page:1 list index out of range pagina sin informacion a: ('1103', 'RECTORIA'), per: 092020, page:2 list index out of range m: 09 a: 2020 d: ('1104', 'SECRETARIA GENERAL') p: ['1', '2', '3'] pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 092020, page:1 list index out of range pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 092020, page:2 list index out of range pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 092020, page:3 list index out of range m: 09 a: 2020 d: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS') p: ['1', '2', '3'] pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 092020, page:1 list index out of range pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 092020, page:2 list index out of range pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 092020, page:3 list index out of range m: 09 a: 2020 d: ('1202', 'SRIA. DE INVESTIGACIÓN CIENTÍFICA Y DESARROLLO TECNOLÓGICO') p: ['1'] pagina sin informacion a: ('1202', 'SRIA. DE INVESTIGACIÓN CIENTÍFICA Y DESARROLLO TECNOLÓGICO'), per: 092020, page:1 list index out of range m: 10 a: 2020 d: ('1103', 'RECTORIA') p: ['1', '2'] pagina sin informacion a: ('1103', 'RECTORIA'), per: 102020, page:1 list index out of range pagina sin informacion a: ('1103', 'RECTORIA'), per: 102020, page:2 list index out of range m: 10 a: 2020 d: ('1104', 'SECRETARIA GENERAL') p: ['1', '2', '3'] pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 102020, page:1 list index out of range pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 102020, page:2 list index out of range pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 102020, page:3 list index out of range m: 10 a: 2020 d: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS') p: ['1', '2', '3'] pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 102020, page:1 list index out of range pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 102020, page:2 list index out of range pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 102020, page:3 list index out of range m: 10 a: 2020 d: ('1202', 'SRIA. DE

INVESTIGACIÓN CIENTÍFICA Y DESARROLLO TECNOLÓGICO') p: ['1'] pagina sin informacion a: ('1202', 'SRIA. DE INVESTIGACIÓN CIENTÍFICA Y DESARROLLO TECNOLÓGICO'), per: 102020, page:1 list index out of range m: 11 a: 2020 d: ('1103', 'RECTORIA') p: ['1', '2'] pagina sin informacion a: ('1103', 'RECTORIA'), per: 112020, page:1 list index out of range pagina sin informacion a: ('1103', 'RECTORIA'), per: 112020, page:2 list index out of range m: 11 a: 2020 d: ('1104', 'SECRETARIA GENERAL') p: ['1', '2', '3'] pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 112020, page:1 list index out of range pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 112020, page:2 list index out of range pagina sin informacion a: ('1104', 'SECRETARIA GENERAL'), per: 112020, page:3 list index out of range m: 11 a: 2020 d: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS') p: ['1', '2', '3'] pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 112020, page:1 list index out of range pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 112020, page:2 list index out of range pagina sin informacion a: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS'), per: 112020, page:3 list index out of range m: 11 a: 2020 d: ('1202', 'SRIA. DE INVESTIGACIÓN CIENTÍFICA Y DESARROLLO TECNOLÓGICO') p: ['1'] pagina sin informacion a: ('1202', 'SRIA. DE INVESTIGACIÓN CIENTÍFICA Y DESARROLLO TECNOLÓGICO'), per: 112020, page:1 list index out of range list index out of range m: 12 a: 2020 d: ('1103', 'RECTORIA') p: [] list index out of range m: 12 a: 2020 d: ('1104', 'SECRETARIA GENERAL') p: [] list index out of range m: 12 a: 2020 d: ('1201', 'DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS') p: [] list index out of range m: 12 a: 2020 d: ('1202', 'SRIA. DE INVESTIGACIÓN CIENTÍFICA Y DESARROLLO TECNOLÓGICO') p: []

```
df = pd.read_csv("uanl.csv")
df["dependencia"] = ["RECTORIA" for i in range(0, obtener_cantidad_de_filas(df))]
df["mes"] = ["10" for i in range(0, obtener_cantidad_de_filas(df))]
df["anio"] = ["2020" for i in range(0, obtener_cantidad_de_filas(df))]
df["Sueldo Neto"] = df["Sueldo Neto"].transform(limpiar_dato_sueldo)
print_tabulate(df)
```

Traceback (most recent call last): File "*home/jhernandez*.pyvenv/37/lib64/python3.7/site-packages/pandas/core/indexes/base.py", line 2891, in get_{loc} return $\text{self.}_{\text{engine.get loc}}(\text{casted}_{\text{key}})$

File “pandas/$_{\text{libs}}$/index.pyx”, line 70, in pandas.$_{\text{libs.index.IndexEngine.get\_loc}}$ File “pandas/$_{\text{libs}}$/index.pyx”, line 101, in pandas.$_{\text{libs.index.IndexEngine.get\_loc}}$ File “pandas/$_{\text{libs}}$/hashtable$_{\text{class_helper.p}}$ line 1675, in pandas.$_{\text{libs.hashtable.PyObjectHashTable.get\_item}}$ File “pandas/$_{\text{libs}}$/hashtable$_{\text{class_helper.pxi}}$”, line 1683, in pandas.$_{\text{libs.hashtable.PyObjectHashTable.get\_item}}$ KeyError: ’Sueldo Neto’

The above exception was the direct cause of the following exception:

Traceback (most recent call last): File “<stdin>”, line 1, in <module> File “/tmp/babel-t856sX/python-DzrQrO”, line 6, in <module> df[“Sueldo Neto”] = df[“Sueldo Neto”].transform(limpiar$_{\text{datosueldo}}$) File “*home/jhernandez*.pyvenv/37/lib64/python3 packages/pandas/core/frame.py”, line 2902, in getitem indexer = self.columns.get$_{\text{loc}}$(key) File “*home/jhernandez*.pyvenv/37/lib64/python3.7/site-packages/pandas/core/indexes/base.py”, line 2893, in get$_{\text{loc}}$ raise KeyError(key) from err KeyError: ’Sueldo Neto’

1.2 wiki

```
def wiki() -> pd.DataFrame:
    soup = get_soup("https://en.wikipedia.org/wiki/List_of_states_of_Mexico")
    list_of_lists = [] # :List
    rows = soup.table.find_all('tr')
    for row in rows[1:]:
        columns = row.find_all('td')
        #  listado_de_valores_en_columnas = []
        #  for column in columns:
        #    listado_de_valores_en_columnas.append(coulmn.text.strip())
        listado_de_valores_en_columnas = [column.text.strip() for column in columns]
        list_of_lists.append(listado_de_valores_en_columnas)

    return pd.DataFrame(list_of_lists, columns=[header.text.strip() for header in  rows


df = wiki()
print_tabulate(df)
df.to_csv("estados.csv", index=False)
```

	State	Official name (except Mexico City): Estado Libre y Soberano de (English: “Free and Sovereign State of”) :	C
0	Aguascalientes	Aguascalientes	
1	Baja California	Baja California	
2	Baja California Sur	Baja California Sur	
3	Campeche	Campeche	
4	Chiapas	Chiapas	
5	Chihuahua	Chihuahua	
6	Mexico City	Ciudad de México	
7	Coahuila1 4	Coahuila de Zaragoza	
8	Colima6	Colima	
9	Durango	Durango	
10	Guanajuato	Guanajuato	
11	Guerrero	Guerrero	
12	Hidalgo	Hidalgo	
13	Jalisco	Jalisco	
14	México	México	
15	Michoacán	Michoacán de Ocampo	
16	Morelos	Morelos	
17	Nayarit	Nayarit	
18	Nuevo León4	Nuevo León	
19	Oaxaca	Oaxaca	
20	Puebla	Puebla	
21	Querétaro	Querétaro de Arteaga	
22	Quintana Roo	Quintana Roo	
23	San Luis Potosí	San Luis Potosí	
24	Sinaloa	Sinaloa	
25	Sonora2	Sonora	
26	Tabasco5	Tabasco	
27	Tamaulipas4	Tamaulipas	
28	Tlaxcala	Tlaxcala	
29	Veracruz	Veracruz deIgnacio de la Llave	
30	Yucatán3	Yucatán	
31	Zacatecas	Zacatecas	

1.3 csv

1.3.1 from file

```
df = pd.read_csv("/home/jhernandez/Sync/FCFMClases/21-1FJ/DataMining/dm_lmv_6.csv")
print_tabulate(df)
```

	No	Matricula	Nombre del Alumno	Oportunidad
0	1	1809913	ARIZPE CURA DANIEL ALEJANDRO	1
1	2	1813847	GARZA GONZALEZ LETICIA STEPHANIE	1
2	3	1687417	GONZALEZ OLIVARES FRANCISCO JAVIER	1
3	4	1735835	GUEVARA MARTINEZ JOSE JUAN	1
4	5	1850325	HERNANDEZ GONZALEZ OMAR	1
5	6	1723927	QUIROZ SANCHEZ PAULINA	1
6	7	1811144	ROBLEDO HERRERA VICTOR MANUEL	1
7	8	1467561	ZAPATA MEDINA GUSTAVO ANTONIO	1
8	9	1798374	GAMEZ BALDERAS ALAN	1
9	10	1743407	MENDEZ MALDONADO MARLON ISRAEL	1
10	11	1884124	PARDO GAYTAN PEDRO RICARDO	1
11	12	1842600	SANCHEZ AVILA ESTEBAN	1
12	13	1803058	SIFUENTES CORTEZ SAMUEL	1

1.3.2 from url

```
df = get_csv_from_url("https://raw.githubusercontent.com/cs109/2014_data/master/countri
print_tabulate(df)
df.to_csv("paises.csv", index=False)
```

	Country	Region
0	Algeria	AFRICA
1	Angola	AFRICA
2	Benin	AFRICA
3	Botswana	AFRICA
4	Burkina	AFRICA
5	Burundi	AFRICA
6	Cameroon	AFRICA
7	Cape Verde	AFRICA
8	Central African Republic	AFRICA
9	Chad	AFRICA
10	Comoros	AFRICA
11	Congo	AFRICA
12	Congo, Democratic Republic of	AFRICA
13	Djibouti	AFRICA
14	Egypt	AFRICA
15	Equatorial Guinea	AFRICA
16	Eritrea	AFRICA
17	Ethiopia	AFRICA
18	Gabon	AFRICA
19	Gambia	AFRICA
20	Ghana	AFRICA
21	Guinea	AFRICA
22	Guinea-Bissau	AFRICA
23	Ivory Coast	AFRICA
24	Kenya	AFRICA
25	Lesotho	AFRICA
26	Liberia	AFRICA
27	Libya	AFRICA
28	Madagascar	AFRICA
29	Malawi	AFRICA
30	Mali	AFRICA
31	Mauritania	AFRICA
32	Mauritius	AFRICA
33	Morocco	AFRICA
34	Mozambique	AFRICA
35	Namibia	AFRICA
36	Niger	AFRICA
37	Nigeria	AFRICA
38	Rwanda	AFRICA
39	Sao Tome and Principe	AFRICA
40	Senegal	AFRICA
41	Seychelles	AFRICA
42	Sierra Leone	AFRICA
43	Somalia	AFRICA
44	South Africa	AFRICA
45	South Sudan	AFRICA
46	Sudan	AFRICA
47	Swaziland	AFRICA